Information Retrieval and Classification based

Approaches for the Sexual Predator Identification

Darnes Vilariño, Esteban Castillo, David Pinto

Iván Olmos, Saul León

*Benemérita Universidad Autónoma de Puebla, Faculty of Computer Science, Puebla, 72000, Mexico*

*{darnes, dpinto, iolmos}@cs.buap.mx, ecjbuap@gmail.com, saul.ls@live.com*

Nowadays there has been a growing number on the use of messaging systems such as chats and instant messaging which provide sexual predators a good platform for sexual purposes. Thus, it becomes very important to tackle the problem of sexual predator identification in order to ameliorate cases of sexual harassment. In this context, the 6th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN’12) has proposed a task named Sexual Predator Identification with the aim of generating a framework in which different teams around the world may compare their approaches solving this particular problem. The goal is to provide an automatic method that permits to detect chat conversations in which one person attempts any erotic or suggestive remarks. The proposed task is subdivided into two sub-tasks:

– Identification of sexual predators. To detect the users considered to be sexual predators by classifying those conversations containing sexual predator behavior.

– Identification of sexual predator chat lines. To detect the specific lines in which the sexual remark is done.

In this paper we present the evaluation of two different approaches with the aim of tackling the task of Sexual Predator Identification of PAN 2012. The first approach uses a dictionary of sexual terms in order to identify those documents associated in some manner with a sexual predator behavior. In order to do so, we use the sexual terms of the dictionary as a query in an information retrieval system, thus, retrieving the documents that best match with the query introduced. The second approach uses the multinomial Naïve Bayes classifier in order to detect sexual predators. The first approach performed better than the second one with low percentages of precision and high values of recall.